МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ СЕМЕНА КУЗНЕЦЯ

ЗАТВЕРДЖЕНО
на засіданні кафедри
кібербезпеки та інформаційних технологій
Протокол № 2 від 29.08.2024 р.

ПОГОДЖЕНО
Проректор з навчально-методичної роботи

Каріна НЕМАШКАЛО

РОЗПОДІЛЕНІ СХОВИЩА ДАНИХ
робоча програма навчальної дисципліни (РПНД)

Галузь знань	**12 "Інформаційні технології"**
Спеціальність	**122 "Комп'ютерні науки"**
Освітній рівень	**другий (магістерський)**
Освітня програма	**" Комп'ютерні науки "**

Статус дисципліни	**обов'язкова**
Мова викладання, навчання та оцінювання	**англійська**

Розробник(и):
доктор технічних наук, професор — підписано КЕП — Володимир АЛЕКСІЄВ

Завідувач кафедри кібербезпеки та інформаційних технологій — Ольга СТАРКОВА

Гарант програми
доктор технічних наук, професор — підписано КЕП — Сергій МІНУХІН

Харків
2024

MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE
SIMON KUZNETS KHARKIV NATIONAL UNIVERSITY OF ECONOMICS

APPROVED
at the meeting of the department
of cybersecurity and
information technologies
Protocol № 2 of 29.08.2024.

AGREED
Vice-rector for educational and methodical work

Karina NEMASHKALO

DISTRIBUTED DATA STORAGE

Program of the course

Field of knowledge	**12 Information technologies**
Specialty	**122 Computer sciences**
Study cycle	**second (master's)**
Study programme	**Computer Sciences**

Course status	**mandatory**
Language	**English**

Developer:
Dr. Sc. (Engineering), prof. — Digitally signed — Volodymyr ALEKSIYEV

Head of Cybersecurity and
Information Technologies
Department — Olha STARKOVA

Head of Study Programme — Digitally signed — Serhii MINUKHIN

Kharkiv
2024

INTRODUCTION

Distributed data storages are the basis for building effective information systems from the stage of data center server configuration to the creation of cloud computing solutions. The principles, methods and technologies for creating, deploying and maintaining distributed data storages are the basis for scaling modern web solutions, web applications and web services. Today, there are many software systems and services for building large-capacity and Internet-accessible data warehouses and data lake systems that provide opportunities to increase the security of information systems, improve their reliability, and provide tools for scaling design solutions. Knowledge and competencies in choosing the architecture for building modern data storage facilities and implementing industrial systems based on them are relevant and necessary in the training of computer science specialists.

The aim of the course "Distributed Data Storage" is to provide a system of theoretical knowledge and acquire practical skills and abilities to apply, configure and administrate systems based on distributed data storage technologies and design appropriate reliable and cost-effective systems for storing large data volumes.

The objectives of the course are: an overview of existing solutions from the single server level, building networked data warehouses (SAN and NAS systems), cluster file and object data storages, and determining the role of decentralized systems in industrial solutions. Along with this, the discipline considers the features of building and scaling relational databases and NoSQL solutions.

The subject of the course is software tools for building distributed data storages.

The object of the course is the processes of deployment, administration and implementation of solutions for data storage and organization of secure access to them.

The learning outcomes and competencies formed by the course are defined in table 1.

Table 1

Learning outcomes and competences formed by the course

Learning outcomes	**Competencies**
LO4.	GC07, SC09.
LO6.	GC02, GC05, GC07, SC01, SC02, SC05, SC09.
LO7.	GC01, GC03, SC01.
LO9.	GC02.
LO10.	GC07, SC09.

End of the table. 1

Learning outcomes	Competencies
LO12.	GC01, GC02, GC03, GC05, GC07, SC02, SC04, SC05, SC06, SC07, SC08, SC09, SC11, SC12.
LO13.	GC02, SC05.
LO17.	SC05.
LO18.	SC09.
LO20.	GC07, SC05, SC09.

where, LO4. Manage work processes in the field of information technology, which are complex, unpredictable and require new strategic approaches.

LO6. Develop a conceptual model of an information or computer system.

LO7. Develop and apply mathematical methods for the analysis of information models.

LO9. Develop algorithmic and software for data analysis (including large data).

LO10. To design architectural solutions of information and computer systems for various purposes.

LO12. Design and support databases and knowledge.

LO13. Assess and ensure the quality of information and computer systems for various purposes.

LO17. Identify and eliminate problem situations during software operation, formulate tasks for its modification or reengineering.

LO18. Collect, formalize, systematize and analyze the needs and requirements for the information or computer system being developed, operated or supported.

LO20. Develop algorithms and software components for computer information systems for high-performance big data processing systems (including distributed and parallel computing) and cloud platform services.

GC01. Ability to abstract thinking, analysis, and synthesis.

GC02. Ability to apply knowledge in practical situations.

GC03. Ability to communicate in the national language both orally and in writing.

GC05. Ability to learn and master modern knowledge.

GC07. Ability to generate new ideas (creativity).

SC01. Awareness of the theoretical foundations of computer science.

SC02. Ability to formalize the subject area of a particular project in the form of an appropriate information model.

SC04. The ability to collect and analyze data (including large data) to ensure the quality of project decision-making.

SC05. Ability to develop, describe, analyze and optimize architectural solutions of information and computer systems for various purposes.

SC06. Ability to apply existing and develop new algorithms for solving problems in the field of computer science.

SC07. Ability to develop software in accordance with the formulated requirements, considering available resources and constraints.

SC08. Ability to develop and implement software development projects, including in unpredictable conditions, with unclear requirements and the need to apply new strategic approaches, use software tools to organize teamwork on the project.

SC09. Ability to develop and administer databases and knowledge bases.

SC11. Ability to initiate, plan and implement the development processes of information and computer systems and software, including its development, analysis, testing, system integration, implementation and support.

SC12. Ability to develop, apply and integrate data processing and analysis technologies in high-performance systems and cloud platforms to ensure efficient use of computing resources of computer systems.

COURSE CONTENT

Content module 1. Distributed file systems and modern database systems.

Topic 1. Distributed data storage technologies for solving the problems of processing large volumes of data.

1.1. Classification of technologies for creating distributed data storages and database systems. Determination of the main areas of study of the discipline.

1.2. Big Data processing and Data Lake technologies.

Identification of the problem of processing large amounts of data. Solutions for building heterogeneous data warehouses based on cloud computing.

1.3. State-of-the-art data center. Virtualization technologies.

Modern technologies of virtualization of the workstation (Desktop) and server (Server) level. Fundamentals of building cloud computing tools. Private cloud. Data storage layer.

1.4. Server virtualization tools.

Live migration technology. Differences between hypervisor-based virtualization systems and container application level.

Topic 2. Distributed file-like storages based on SAN and NAS technologies. Cloud folders.

2.1. Building data storage at the separate server.

Features of file systems. LVM technology. Use of drives that are combined into a RAID array.

2.2. Technologies of network data storages.

Application of SAN and NAS solutions by the example of TrueNAS technologies.

2.3. Solutions for building cloud folders.

Practical experience with ownCloud and Nextcloud systems.

Topic 3. Object data stores. Clustered file systems.

3.1. Clustered file systems.

Building a cluster for storing data by the example of the GoogleFS concept. MapReduce algorithm.

3.2. Ceph Distributed Data Storage.

Provision of data storage on example of the Ceph cluster deployment.

Topic 4. Scaling data storage systems and creating knowledge bases on relational DBMS platform.

4.1. Features of OLAP (online analytical processing) technology.

Application of relational databases as a platform for building knowledge bases for decision-making systems.

4.2. Database replication.

Solving data replication problems using MySQL as an example.

4.3. Building a relational database cluster.

PostgreSQL cluster building tools. Galera Cluster solutions for MariaDB or MySQL.

Topic 5. Non-relational database technologies.

5.1. Features of the key-value database.

Distributed storage, by the example of Redis technology, for solving the problems of building scalable web applications.

5.2. MongoDB database.

Features of application and scaling of solutions built based on a non-relational database.

5.3. Scalable and reliable data warehouses based on Hadoop and Cassandra technologies.

Building solutions for storing and processing big data.

Content module 2. Application and features of design solutions based on distributed data storages.

Topic 6. Monitoring the state of distributed computing systems and data warehouses based on open-source software.

6.1. Monitoring by Nagios, Zabbix, etc. tools.

Features of the use of common multi-agent monitoring systems.

6.1. Prometheus monitoring and Grafana data visualization.

Features of the use of monitoring systems, which are built on a modular principle.

Topic 7. Features of the development of web applications and web services using distributed data storage technologies.

7 .1. The CAP theorem (Brewer's theorem).

A combination of data consistency, availability, and separation tolerance in distributed systems.

7.2. Features of scaling web applications.

Understanding the role of data warehouses in solving the problems of building web applications. Features of the architecture of modern web applications, for example: the use of microservices, the use of message chains, etc.

7.3. Centralized and decentralized systems.

Features of building decentralized data storage systems. Technologies of torrent networks. Blockchain technology.

Topic 8. Distributed storage technologies in cloud computing.

Features of the implementation of the "data storages as a service" paradigm in cloud solutions: Amazon Web Services, Microsoft Azure, Google Cloud and Oracle Cloud Infrastructure.

Topic 9. Prospects for the development of distributed data storage systems and technologies.

Overview of current trends and prospects for the development of technologies and complex solutions for the construction and application of distributed data storages.

The list of laboratory classes / tasks for the course is given in Table. 2.

Table 2

List of laboratory classes / tasks

Title of the topic and/or task	Content
Topic 2. Task 1.	Deploy TrueNAS or equivalent in a virtualization environment. Open FTP access to save and download files.
Topic 2. Task 2.	Investigate the operation of ownCloud or Nextcloud tools.
Topic 3. Task 3.	Deploy a Ceph cluster (three nodes) in a virtualization environment. Investigate the operation of the relevant solution.
Topic 4. Task 4.	In a virtualization environment, investigate the features of configuring MySQL replication (relational database).
Topic 4. Task 5.	In a virtualization environment, investigate the features of deploying a relational database cluster using Galera Cluster or analogue technology.
Topic 5. Task 6.	Deploy three MongoDB nodes and use it to investigate the MapReduce algorithm (NoSQL solution).

The list of self-studies work in the course is given in Table. 3.

Table 3

List of self-studies work

Title of the topic and/or task	Content
Topic 1. Task 1.	According to the Internet sources, consider architectural solutions for building Data Lake.
Topic 2. Task 2.	Deploy a virtual machine based on a modern distribution, for example, Ubuntu Server. Determine the feasibility of using LVM technology. Explore software-RAID capabilities. Deploy an Apache web server and provide users with the ability to upload web application files using FTP technology. Perform a security study of the relevant solution.
Topic 5. Task 3.	Explore the deployment and application of Redis or equivalent solutions.
Topic 5. Task 4.	According to the Internet, compare the features of Hadoop and Cassandra technologies.
Topic 6. Task 5.	Deploy (if technically possible) and investigate Zabbix monitoring tools.
Topic 6. Task 6.	Deploy (if technically possible) and perform research on Prometheus monitoring and Grafana data visualization tools.
Topic 7. Task 7.	According to the Internet, consider architectural solutions for building decentralized data storage systems.
Topic 8. Task 8.	Based on the materials of the Internet, perform research and compare the implementation of services that use distributed data storage in cloud computing solutions.
Topic 9. Task 9.	Based on the materials of the Internet, to study the trends in the development of distributed data storage technologies. Essay writing.

The number of hours of lectures, laboratory classes and hours of independent work is given in the work plan (technological map) for the discipline.

TEACHING METHODS

In the process of teaching the discipline for the acquisition of certain learning outcomes, the activation of the educational process, the use of such teaching methods as the following is provided:

Verbal (lecture (Topic 2, 3, 4, 5, 6, 8), problem lecture (Topic 1, 7, 9)).

Visual (demonstration (Topic 2, 3, 5, 6)).

Practical (laboratory work (Topic 2, 3, 4, 5), essay (Topic 9), case method (Topic 8)).

FORMS AND METHODS OF ASSESSMENT

The university uses a 100-point cumulative system for assessing the learning outcomes of higher education applicants.

Current control is carried out during lectures, laboratory classes and is aimed at checking the level of readiness of the student to perform a specific job and is

evaluated by the amount of points scored: for courses with a form of semester control as grading: maximum amount is 100 points; minimum amount required is 60 points.

The final control includes current control and assessment of the student.

Semester control is carried out in the form of a grading.

The final grade in the course is determined: for disciplines with a form of grading, the final grade is the amount of all points received during the current control.

During the teaching of the course, the following control measures are used:

Current control: performance and defense of laboratory workshop (6 works with 10 points each), written tests (4 works with 10 points each).

Semester control: Grading.

More detailed information on the assessment system is provided in technological card of the course.

RECOMMENDED LITERATURE

Main

1. Azure Strategy and Implementation Guide, Fourth Edition / Jack Lee, Greg Leonardo, Jason Milgram, and David Rendón, Packt Publishing, 2021. – 237 p. [Electronic resource]. – Access mode: https://info.microsoft.com/ww-landing-azure-strategy-and-implementation-guide.html.

2. Azure for Architects Third Edition / Ritesh Modi, Jack Lee, and Rithin Skaria, Packt Publishing, 2020. – 701 p. [Electronic resource]. – Access mode: https://info.microsoft.com/ww-landing-azure-for-architects.html.

3. Professional Azure SQL Managed Database Administration – Third Edition / Ahmad Osama and Shashikant Shakya, Packt Publishing, 2021. – 725 p. [Electronic resource]. – Access mode: https://info.microsoft.com/ww-landing-professional-azure-sql-database-administration-packt.html.

4. IBM Storage Ceph Concepts and Architecture Guide / Vasfi Gucer, Jussi Lehtinen, Jean-Charles (JC) Lopez, Christopher Maestas, Franck Malterre, Suha Ondokuzmayis, Daniel Parkes and John Shubeck, IBM Corp. 2024. – 172 p. https://www.redbooks.ibm.com/abstracts/redp5721.html.

5. Oracle Database Database Concepts, 23ai / Mark Doran, Lance Ashdown, Donna Keesling, Tom Kyte, Oracle and/or its affiliates., 2024. – 733 p. [Electronic resource]. – Access mode: https://docs.oracle.com/en/database/oracle/oracle-database/23/cncpt/index.html.

6. Building Big Data and Analytics Solutions in the Cloud / Wei-Dong Zhu, Manav Gupta, Ven Kumar, Sujatha Perepa, Arvind Sathi and Craig Statchuk, IBM Corp. 2014. – 114 p. [Electronic resource]. – Access mode: https://www.redbooks.ibm.com/abstracts/redp5085.html.

Additional

7. Initial Server Setup with Ubuntu [Electronic resource] / Jamon Camisso and Anish Singh Walia. DigitalOcean, 2022. – Access mode: https://www.digitalocean.com/community/tutorials/initial-server-setup-with-ubuntu.

8. Jamon Camisso. Sysadmin eBook: Making Servers Work – DigitalOcean, 2020. – 281 p. [Electronic resource]. – Access mode: https://www.digitalocean.com/community/books/sysadmin-ebook-making-servers-work.

9. M. van Steen and A.S. Tanenbaum, Distributed Systems, 4th ed., distributed-systems.net, 2023. – 685 p. [Electronic resource]. – Access mode: https://www.distributed-systems.net/index.php/books/ds4/.

10. Kravchenko, P. Blockchain and decentralized systems. : in three volumes. Vol. 1 / P. Kravchenko, B. Skriabin, O. Dubinina. – Kharkiv, 2018. – 416 p.: – Access mode: https://books.distributedlab.com/books-en/#extras.

11. Kravchenko, P. Blockchain and decentralized systems : in three volumes. V.2 / P. Kravchenko, B. Skriabin, O. Kurbatov, O. Dubinina. – Kharkiv, 2019. – 396 p. – Access mode: https://books.distributedlab.com/books-en/#extras.

12. Kravchenko, P. Blockchain and decentralized systems : in three volumes. V.3 / P. Kravchenko, B. Skriabin, O. Kurbatov, O. Dubinina. – Kharkiv : 2020. – 298 p. – Access mode: https://books.distributedlab.com/books-en/#extras.

13. Martovytskyi V. Technology for monitoring the functioning state of distributed computer systems / V. Martovytskyi, Y. Koltun, D. Holubnychyi et al. // Системи управління, навігації та зв'язку : зб. наук. пр. – 2022. – Вип. 1 (67). – С. 75-80. [Electronic resource]. – Access mode: http://www.repository.hneu.edu.ua/handle/123456789/27369.

14. Minukhin S. Performance study of the DTU model for relational databases on the Azure platform / S. Minukhin // Сучасний стан наукових досліджень та технологій в промисловості. – 2022. – № 1 (19). – С. 27-39. [Electronic resource]. – Access mode: http://repository.hneu.edu.ua/handle/123456789/27739.

Information resources

15. TrueNAS Documentation Hub [Electronic resource]. –Access mode: https://www.truenas.com/docs/core/13.0/gettingstarted/

16. Nextcloud Server Administration Guide [Electronic resource]. –Access mode: https://docs.nextcloud.com/server/latest/admin_manual/

17. The Ceph Documentation [Electronic resource]. –Access mode: https://docs.ceph.com/en/latest/

18. MySQL Documentation [Electronic resource]. –Access mode: https://dev.mysql.com/doc/

19. MariaDB Server documentation [Electronic resource]. –Access mode: https://mariadb.org/documentation/

20. What is MariaDB Galera Cluster? [Electronic resource]. –Access mode: https://mariadb.com/kb/en/what-is-mariadb-galera-cluster/

21. MongoDB Documentation [Electronic resource]. –Access mode: https://www.mongodb.com/docs/
22. Zabbix Documentation [Electronic resource]. –Access mode: https://www.zabbix.com/manuals
23. What is Prometheus? [Electronic resource]. –Access mode: https://prometheus.io/docs/introduction/overview/
24. Apache Cassandra documentation [Electronic resource]. –Access mode: https://cassandra.apache.org/doc/latest/
25. Apache Hadoop Documentation [Electronic resource]. –Access mode: https://hadoop.apache.org/docs/current/
26. Site of personal learning systems of Simon Kuznets Kharkiv National University of Economics according to the course of "Distributed data storages" [Electronic resource]. –Access mode: https://pns.hneu.edu.ua/course/view.php?id=10295.